A SURVEY ON MACHINE LEARNING APPROACH TO REDUCE DIMENSIONALITY SPACE IN LARGE DATASETS USING PCA AND LDA

Jessin Shah P A[1], Dr. G Kiruthiga[2]

[1] *PG Scholar, Department of Computer Science and Engineering, IES College of Engineering, Kerala, India*
[2] *Head of the Department, Department of Computer Science and Engineering, IES College of Engineering, Kerala, India*

ABSTRACT

Due to digitization several sectors like health, web organizations, production, generate huge volume of data. To uncover the patterns among the attributes of this data machine learning algorithms can be used and can make predictions that can be used by the medical practitioners and people at different managerial level to make decisions. All attributes gathered might not be contribute to the prediction and by removing irrelevant attributes reduces the burden on machine learning. There are several techniques to reduce the dimension of data gathered. These dimensional reduction techniques help to reduce the dimension and thereby only features attributes can be used to model the system. This increases the performance of the model. The traditional dimensionality reduction techniques mostly used is Principal Component Analysis and Linear Discriminant Analysis In this paper, different dimensionality reduction techniques and their result on the performance of the machine learning techniques is studied.

Keyword : - *PCA, LDA, dimensionality reduction tehniques*

1. INTRODUCTION

Data is growing at a fast rate and the demand for its storage space is also high. Depending on type of process, data collected from different sources create huge datasets of multidimensional. Analysis of high dimensional dataset is difficult. In such situations dimension reduction plays an important role, where reducing number of arbitrary attributes under considerations[1]. This is the process of reducing some features or columns in dataset without losing much of vital aspects of the primary dataset. Different methods used in dimensionality reduction are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Generalized Discriminant Analysis (GDA). Dimensionality reduction can be linear or non linear based on the methods used. Some advantages of dimensionality reduction include eliminate irrelevant features, data compression, reduce computation time.

Oldest and best known method of multivariate data analysis is the principal component analysis. It was first invented by Karl Pearson in 1901 and later independently developed by Harold Hotelling in 1933. Principal Component Analysis is a technique which uses underlying mathematical principles to transform a number of possibly correlated variables into smaller number of variables called principal components. Its most common use is as the first step in trying to analyze large datasets. PCA performs a linear mapping of data to a low-dimensional space. Variance of the data in low-dimensional representation is maximized. In practical, covariance matrix of data

is constructed and eigen vectors on this matrix are computed. Eigen vectors corresponds to largest eigen values are called principal components. These are used to reconstruct a large fraction of variance of original data. First few eign vectors contribute the vast majority of system's energy, especially in low-dimensional systems.

For more than two classes, LDA is preferred. It consists of statistical properties of data, that is mean and variance, is calculated for each class. For multiple variables, same poperties is calculated over multivariate Gaussian, that is mean and covariance matrix. These are given to LDA equation to make predictions.

In GDA, nonlinear discriminant analysis using kernel function operator is performed. This find projection for the features into lower dimensional space, by maximizing the ratio of between-class scatter to within-class scatter.

2. LITERATURE SURVEY

Two popular dimensionality reduction techniques are namely Linear Discriminant Analysis (LDA) and PCA. These techniques are investigated on widely used ML algorithms namely Decision Tree, Naïve Bayes, Random Forest and Support Vector Machine. A novel architecture of implementing dimensionality reduction technique using PCA is divided into five phases that presents a hybrid method of machine learning for reducing dimensional space in large datasets. The five phases include data collection, Data pre-processing, data integration, data transformation, datamining and knowledge discovery, applying ML techniques together with PCA. To verify the correctness of the method, a case study with complex dataset, an epileptic seizure recognition database is performed. The results are very promising. First merge all datasets as huge one, after applying ETL process, PCA for dimensionality reduction is applied. Random Forest shows better accuracy than other ML algorithms when experimented with epileptic seizure recognition dataset [1]. Fig. 1 shows PCA representation for dimensionality reduction.

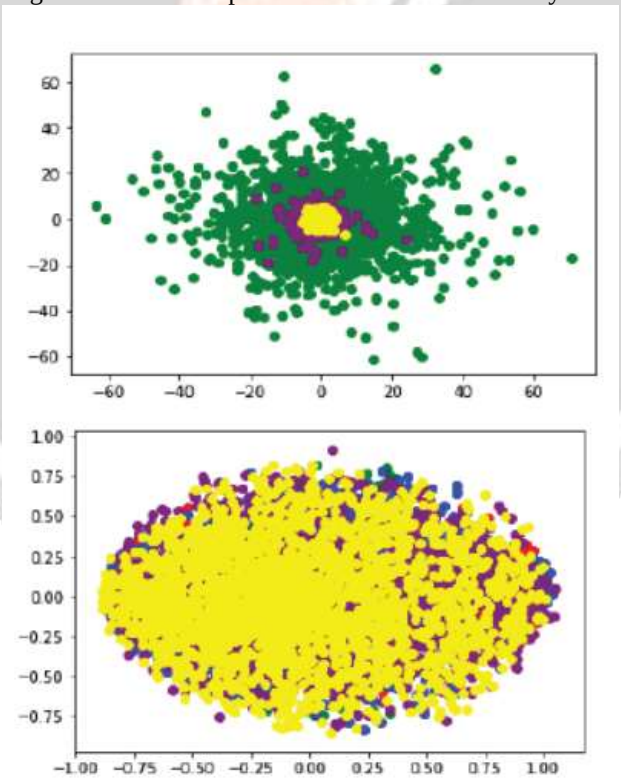

Fig-1: PCA representation for dimension reduction

Feature engineering can be applied to improve the quality of dataset. In this case to extract the important attributes, dimensionality reduction techniques, PCA and LDA is applied individually. Extracted features is fed to the abovementioned ML algorithms. Then the performance of ML algorithms without dimensionality reduction is compared with the performance of ML algorithms with the application of PCA and LDA using several performance metrics. Fig 2 shows performance evaluation of classifiers without dimensionality reduction.

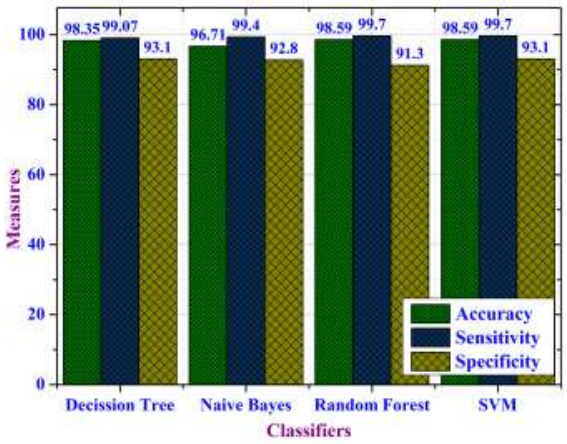

Fig -2: Performance evaluation of classifiers without dimensionality reduction

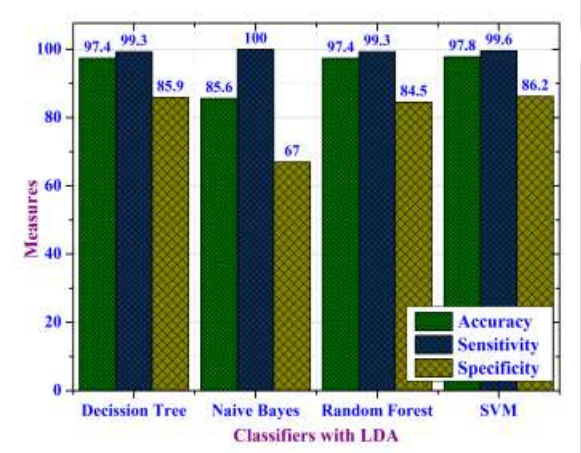

Fig -3: performance evaluation of classifiers using LDA for dimensionality reduction

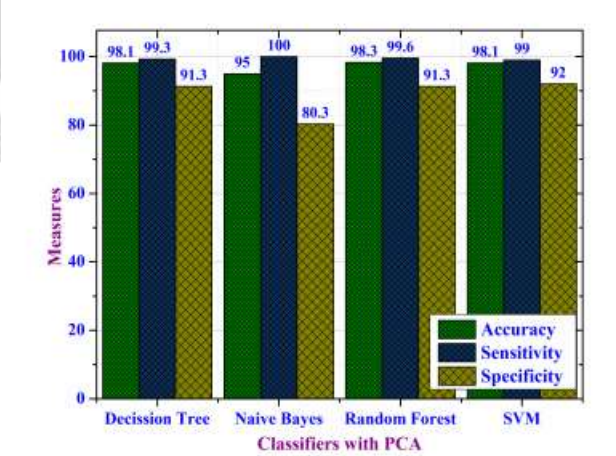

Fig -4: Performance evaluation of classifiers using PCA for dimensionality reduction

Then analyse in detail about the impact of feature engineering and dimensionality reduction techniques on the performance of ML algorithms. From the results obtained, it is observed that performance of classifiers with PCA is better than that with LDA. Decision Tree and Random Forest outperforms the other algorithms without using the dimensionality reduction as well as with both PDA and LDA[2].

Adaptive hybrid feature selection based classifier ensemble (AHFSE) for epileptic seizure classification, creates new sample subsets in every bootstrap and is used within bagging process to obtain new compact sample subsets. In this framework AHFSE algorithm selects the representative features by means of rank aggregation to avoid redundancy and to improve performance of the classifier. Voting method is used to obtain the final detection classification results. Highlights of this experiment are using AHFSE within bagging process improves specificity, sensitivity or accuracy compared to traditional method and also ths helped to obtain the optimum feature selection. The algorithm needs longer time to obtain compact sample subsets [3].

A new framework for multi-label learning with a large number of class labels and features namely MLL-FLSDR. Both feature space and label space are reduced to low dimensional spaces respectively, in which local structure of data points is utilized to constrain the geographical structure on both the learned low dimensional space and guarantee the qualities of them[4].

Neighbors-based distance measures distance of two examples through their neighbors [5]. This can be applied to classical dimensionality reduction methods, whether linear or nonlinear, and achieve substantial performance. An algorithm named DRWPCA reduces dimension by analysing the correlation between the dimensions and therefore the physical meaning of the original dataset is retained. This utilized optimization algorithm to get most suitable parameters, so that algorithm performance is optimal. This shows better dimensionality reduction ability and higher accuracy than PCA [6].

The advantage of an unsupervised multiple layered sparse autoencoder model is that its optimization goal it reduces the reconstruction error, and the resulting low-dimensional feature can be reconstructed to the original dataset as much as possible. Therefore, the reduction of datasets is effective. The relationship among the reconstructed data, the number of iterations, and the number of hidden variables is first explored. Second, proven the dimensionality reduction ability of the sparse autoencoder. On publicly available datasets, several classical feature representation methods are compared with the sparse autoencoder, and the corresponding low-dimensional representations are placed into different supervised classifiers and the classification performances reported. Lastly, the parametric sensitivity is shown by adjusting the parameters that might influence the classification performance. The more efficient and reliable is sparse 8 autoencoder. Unsupervised dimensionality reduction of several datasets was undertaken using the SAE. Low-dimensional features were extracted and classified into predictive labels without labels. Under the appropriate number of iterations and parameters, the ability of SAE to compress and reconstruct images has a good effect, and more adaptive and robust than other dimensionality reduction algorithms. In data classification, assume that the elements in the samples are independent of each other. Improvement in the classification standard was not explored further in the present study [7].

Several computer-aided diagnosis (CAD) systems based on Deep Neural Networks improve the diagnosis for AD and PD and outperform those based on classical classifiers. In order to address the small sample size problem, two dimensionality reduction algorithms based on Principal Component Analysis and Non-Negative Matrix Factorization (NNMF), respectively are evaluated. The performance of CAD systems is assessed using 4 datasets with neuroimaging data of different modalities. For three databases, the proposed systems based on DNN along with PCA dimensionality reduction techniques achieved the highest accuracy. Compared with other classifiers, the accuracy measure obtained by DNN-based systems is more fluctuating. This is especially significant for dataset 1 where rates oscillate between 78 and 90 for NB, DT and SVM but are between 70 and 100 for DNN. The reason can be due to DNN requires larger training samples and during training, has more parameters that should be optimized. In spite of that, the accuracy rate obtained by the DNN-based system, compared to other systems, in the whole CV procedure for dataset 1 is higher. High accuracy rates obtained in this work are partly due to the dimensionality reduction carried out by the algorithms based on PCA and NNMF, which made the classification problem affordable and allowed them to develop full-automatic systems that did not require initialization steps based on previous knowledge. All the data included in this dataset correspond to subjects with early-stage diseases. In classification studies related with clinical diagnosis there often exists an error due to possible labelling errors. An important matter in the development of CAD systems is parameter optimization. For more sophisticated system, more parameters should be optimized. The parameter optimization requires additional data and if data are limited, then simple

systems often perform better than sophisticated ones. The proposed systems also have this problem. For example, the number of features resulting from the dimensionality reduction procedure (parameter k in NNMF or number of principal components in PCA) can be optimized. Similarly, classification parameters such as the cost parameter of the SVM classifier, C, or the number of units of the DNN algorithm can also be optimized. However, this optimization would require additional data and the number of samples in most of the neuroimaging studies (including the ones used in this work) is 10 reduced. For this reason they decided to keep these parameters to their common/default values. In their opinion, the value of the results is not affected by this decision because all the systems were evaluated under the same conditions, so they can be fairly compared [8].

Comparing with Latent Dirichlet allocation (LDA) with probabilistic latent semantic indexing (pLSI) as a dimensionality reduction method for document and investigate their effectiveness in document clustering by using real-world document sets. For clustering of documents, use a method based on multinomial mixture, which is known as an efficient framework for text mining. F-measure is used to evaluate clustering results, i.e., harmonic mean of precision and recall. Japanese and Korean Web articles used for evaluation and regard the category assigned to each Web article as the ground truth for the evaluation of clustering results. The dimensionality reduction via LDA and pLSI results in document clusters of almost the same quality as those obtained by using original feature vectors. Therefore, they can reduce the vector dimension without degrading cluster quality. However, there is no meaningful difference between LDA and pLSI. The results suggests that at least for dimensionality reduction in document clustering LDA does not replace pLSI. LDA and pLSI used to reduce the dimension of document feature vectors which are originally of dimension equal to the number of vocabularies. A clustering based on multinomial mixture for the set of the feature vectors of original dimension and for the set of the vectors of reduced dimension is conducted. They also compare LDA and pLSI with random projection. Further, both LDA and pLSI are more effective than random projection, the baseline method. LDA can reduce the dimension of document feature vectors without degrading the quality of document clusters. Further, LDA is far superior to random projection. However, their experiment tells no significant difference between LDA and pLSI. While considering the issue of computational cost, no good reason to promote LDA beside pLSI for dimensionality reduction in document clustering [9].

In Multi-label feature selection method, labels categorizes into two groups: independent labels and dependent labels[10]. By proposing a new feature relevance term, that is, the conditional mutual information between candidate features and each label has given other labels, analyzes the differences between independent labels and dependent labels. Two approaches for learning such label-dependent rules are bootstrapped stacking approach which can be built on top of a conventional rule learning algorithm and the second approach is by adapting the commonly used separate-and-conquer algorithm for learning multi-label rules [11].

A novel method for subset selection of optimal functions is hybrid PCA where relevant features are obtained from total features and tested through Adaptive Boosting algorithm integrated with Support vector Machine. Feature selection with PCA as pre-processing step proven to be effective in computational time and accuracy. This method help to make medical decisions more efficiently and quickly as this proved to be very effective to improve classification results [12].

PCA and K-means algorithm integrated to predict diabetes. PCA applied as first step for dimensional reduction and then K- means applied to cluster the data. The combination of these two techniques provides higher accuracy rate [13].

PCA, PPCA,EM-PCA,GHA, APEX are five PCA algorithms in dimensionality reduction for clustering. Based on brain tumour images, performance of EM-PCA and PPCA work effectively with clustering algorithms. In this, first PCA is applied to MRI images of different sizes, then clustering is applied using K-means and FCM. This leads to higher performance rate [14].

A PCA-firefly based Deep Learning Model for early detection of diabetic retinopathy. PCA- Firefly algorithm selects optimal features and the Deep Neural Network classify the diabetes retinopathy dataset. This yielded a good classification results than other ML approaches considered [15].

3. CONCLUSIONS

With high dimensional datasets, there is an increase in number of features. If more features than observations are present, then it will overfit the model. The result thus obtained will be out of sample performance. Too many features would make it harder for clustering of observations. Also too many features confuse certain machine learning algorithms. Dimensionality reduction is a solution to such problems. Fewer parameters correspond to fewer parameters in machine learning model. This is referred to as degrees of freedom. Dimensionality reduction is performed prior to data modelling. It is performed after data cleaning and data scaling and before training a predictive model. There are many techniques for dimensionality reduction.

In this paper, a survey on different dimension reduction techniques and their effect on the performance of the machine learning algorithms is presented. This survey covers new methods for the dimensionality reduction and effect on various ML algorithms and a brief description of dimensionality reduction. The details of different dimensionality reduction techniques along with their performance in different datasets is provided. Traditional approach of dimensionality reduction techniques, namely PCA and LDA are applied in most of the experiments. PCA is the oldest and most efficient dimensionality reduction technique. There are several other approaches for dimensionality reductions. But as a pre-processing steps, PCA is commonly and widely used one. ML algorithms using PCA shows better accuracy and performance.

Conclusion related your research work Conclusion related your research work Conclusion related your research work Conclusion related your research work Conclusion related your research work Conclusion related your research work Conclusion related your research work Conclusion related your research work Conclusion related your research work Conclusion related your research work

4. ACKNOWLEDGEMENT

This paper and the research behind it would not have been possible without the exceptional support of my supervisor, G Kiruthiga. I am also grateful for the insightful comments offered by my colleagues allowed me to continue my research with the book much longer than I could have hoped. Finally I would like to thank my college and university for the unconditional support for completing this work.

5. REFERENCES

[1] Rafael Muñoz Terol, Alejandro Reina Reina, Saber Ziaei, And David Gil "A Machine Learning Approach to Reduce Dimensional Space in Large Datasets"
[2] G. Thippa Reddy, M. Praveen Kumar Reddy, Kuruva Lakshmanna, Rajesh Kaluri, Dharmendra Singh Rajput, Gautam Srivastava (Senior Member, Ieee), And Thar Baker "Analysis Of Dimensionality Reduction Techniques On Big Data"
[3] Farrikh Alzami , Juan Tang , Zhiwen Yu, Si Wu , C. L. Philip Chen, Jane You , And Jun Zhang, "Adaptive Hybrid Feature Selection-Based Classifier Ensemble for Epileptic Seizure Classification".
[4] Jun Huang, Pingzhao Zhang, Huiyi Zhang, Guorong Li, And Haowei Rui, "Multi-Label Learning via Feature and Label Space Dimension Reduction"
[5] Hui Tian, Long Lan, Xiang Zhang, And Zhigang Luo, "Neighbors-Based Graph Construction for Dimensionality Reduction"
[6] Rui Zhang, Tao Du , And Shouning Qu , "A Principal Component Analysis Algorithm Based on Dimension Reduction Window"
[7] Jianran Liu, Chan Li, And Wenyuan Yang, "Supervised Learning via Unsupervised Sparse Autoencoder"

[8] F. Segovia, J. M. Górriz, J. Ramírez, F. J. Martinez-Murcia and M. García-Pérez, "Using deep neural networks along with dimensionality reduction techniques to assist the diagnosis of neurodegenerative disorders"
[9] Tomonari Masada, Senya Kiyasu, and Sueharu Miyahara, "Comparing LDA with pLSI as a Dimensionality Reduction Method in Document Clustering"
[10] P. Zhang, G. Liu, and W. Gao, ''Distinguishing two types of labels for multi-label feature selection,'' Pattern Recognit., vol. 95, pp. 72–82, Nov. 2019
[11] E. L. Mencía and F. Janssen, ''Learning rules for multi-label classification: A stacking and a separate-and-conquer approach,'' Mach. Learn., vol. 105, no. 1, pp. 77–126, Oct. 2016.
[12] Y. Zhang and Z. Zhao, ''Fetal state assessment based on cardiotocography parameters using PCA and AdaBoost,'' in Proc. 10th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI), Oct. 2017, pp. 1–6.
[13] C. Zhu, C. U. Idemudia, and W. Feng, ''Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques,'' Informat. Med. Unlocked, vol. 17, 2019, Art. no. 100179.
[14] I. E. Kaya, A. Ç. Pehlivanlı, E. G. Sekizkardeş, and T. Ibrikci, ''PCA based clustering for brain tumor segmentation of T1w MRI images,'' Comput. Methods Programs Biomed., vol. 140, pp. 19–28, Mar. 2017.
[15] T. R. Gadekallu, N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, I.-H. Ra, and M. Alazab, ''Early detection of diabetic retinopathy using PCA-firefly based deep learning model,'' Electronics, vol. 9, no. 2, p. 274, 2020